Syst. Biol. 57(3):503–507, 2008

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150802164587

Hemiplasy: A New Term in the Lexicon of Phylogenetics

JOHN C. AVISE[1] AND TERENCE J. ROBINSON[2]

[1]*Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697, USA; E-mail, javise@uci.edu*
[2]*Evolutionary Genomics Group, Department of Botany and Zoology, University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa*

Homoplasy (trait similarity due to evolutionary convergence, parallelism, or character reversals) is a well-appreciated form of phylogenetic noise that systematists strive to identify and avoid when reconstructing species phylogenies. However, another source of phylogenetic "noise" is often neglected: the idiosyncratic sorting of gene-tree lineages into descendant taxa from character-state polymorphisms retained across successive nodes in a species tree. Here we introduce a term (hemiplasy) that formalizes a category of outcomes that can emerge from this evolutionary lineage-sorting phenomenon, and we make a case for why a wider recognition of hemiplasy (and attempts to ameliorate its complications) can play an important role in phylogenetics.

The word *homoplasy*, meaning shaped (-plasy) in the same (homo-) way, refers to any trait correspondence or similarity not due to common ancestry. A central challenge in phylogenetic reconstruction is thus to distinguish the phylogenetic noise of homoplasy from the phylogenetic signal of homology (similarity in biological features due directly to shared ancestry). However, homology itself bears a subtle relationship to phylogeny, as emphasized by Willi Hennig (1950) more than a half-century ago. Hennig introduced the critical distinction between shared ancestral homology (symplesiomorphic similarity) and shared derived homology (synapomorphic similarity), noting that only the latter is indicative of monophyly within an organismal phylogeny. Hennig's cladistic insights fostered a fundamental revolution in phylogenetic principles and methodologies.

The molecular revolution in biology that began at about that same time added further nuances to the homology concept. For example, DNA sequence homology in a multigene family can be due either to paralogy (similarity tracing to a gene duplication event) or to orthology (similarity tracing to an allelic separation within a particular locus). Orthology and paralogy are both genuine forms of genetic homology, but a failure to distinguish them in comparisons of DNA sequences can lead to errors in phylogenetic reconstruction.

Phylogenetic jargon is already extensive but also important because words such as *homoplasy*, *synapomorphy*, and *orthology* capture and convey sophisticated evolutionary concepts that otherwise might remain opaque or underappreciated. In this spirit, here we formally define a new term—*hemiplasy*—for how the well-known phenomenon of idiosyncratic lineage sorting can lead to fundamental discordances between gene trees and organismal (species) trees. As will be described, hemiplasy is a bona fide form of homology (allelic orthology in this case) that nonetheless can give the illusion of homoplasy in an organismal tree. No other word or simple phrase currently exists to encapsulate the phenomenon that we will define under the suggested term.

CONCEPTUAL BACKGROUND

The nature of Mendelian heredity in sexually reproducing taxa ensures that alleles at unlinked loci transmit through an organismal pedigree via noncoincidental genealogical pathways across multiple generations. Thus, both within and among related species, the true topologies of gene trees inevitably differ somewhat from locus to unlinked locus (Ball et al., 1990). Furthermore, gene genealogies can in principle differ in basic topology from the overall population tree or species tree of which they are a part, if for no other reason than stochastic lineage sorting across successive evolutionary nodes in an organismal phylogeny. These concepts and their corollaries have been available for more than two decades (Hudson, 1983; Tajima, 1983; Takahata and Nei, 1985; Neigel and Avise, 1986), and they are encapsulated

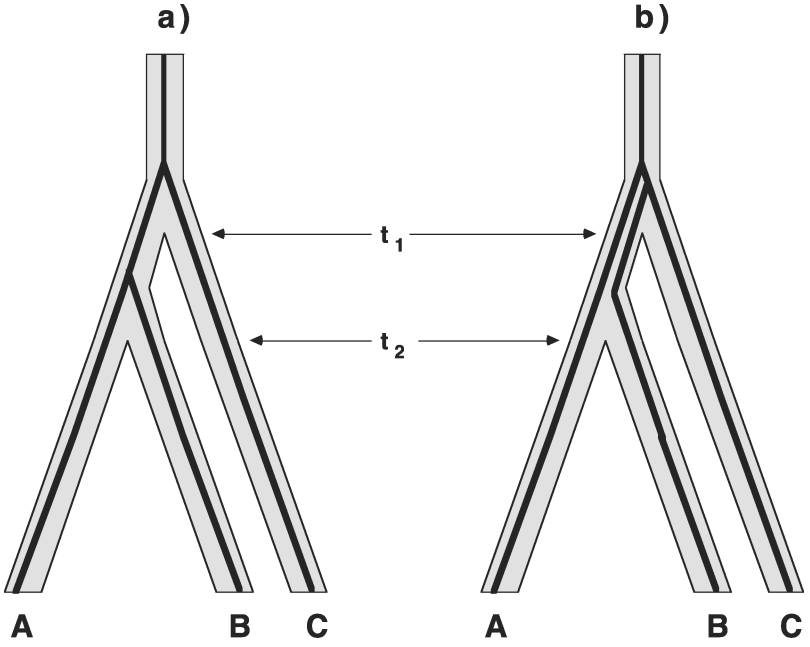

FIGURE 1. Gene genealogies within a three-species phylogeny (after Avise, 2004). Shown are two possible gene trees (thin dark lines), with the gene tree in (b) showing a qualitative topological discordance with the species tree.

today in the widely appreciated distinction between gene trees and organismal trees (Doyle, 1992; Maddison, 1997; Nichols, 2001). Nevertheless, as noted by Liu and Pearl (2007), "the current molecular phylogenetic paradigm still reconstructs gene trees to represent the species tree."

Nei (1987:401–403) summarized the theoretical probability of qualitative discordance between a gene tree and an organismal tree for the simple case of three related species (for more complex situations, see Rosenberg, 2002 and Degnan and Rosenberg, 2006). For selectively neutral alleles, the probability of the topological discordance illustrated in Fig. 1b is given by $(2/3)e^{-T/2N}$, where T is the number of generations between the first and second speciation events ($T = t_1 - t_2$) and N is the effective population size (Fig. 1). This probability can also be interpreted as the percentage of unlinked neutral loci expected to show topological disagreement with the species tree. The formula supports general intuition by showing that a gene tree is most likely to "misrepresent" the topological structure of a species phylogeny when internodal times are short relative to effective population sizes. For example, the discordance probability is approximately 50% when $T/2N = 0.3$, but it is infinitesimally small when $T/2N = 100$.

This type of qualitative discordance between the branching topology of a gene tree and a species tree can also be interpreted to reflect the retention of a polymorphism across successive nodes in a species tree, followed by lineage sorting and idiosyncratic fixation of alternative character states in the descendant species (Fig. 2). Note that allele "b" in Fig. 2 is a derived character state and that it is shared by two descendant taxa (B and C) that nonetheless do not constitute a clade at the organismal level. In other words, character state "b" is a clade-defining synapomorphy, but the monophyletic assemblage that it earmarks is within the gene tree per se rather than at the composite species level.

The probability of topological discordance between a gene tree and a species tree can also reflect additional factors that impact the ratio of $T/2N$. For example, balancing selection can maintain a genetic polymorphism for long periods of time, in effect making $T/2N$ smaller and thereby increasing the probability of an eventual discordance between a species tree and the particular gene tree whose alleles are under selection. Conversely, a genetic polymorphism experiencing underdominant selection, or one whose alleles undergo positive selective sweeps, tends to be transient in a species and thereby is less likely to eventuate in a gene-tree/species-tree incongruence.

Disparities between the topologies of gene trees and species trees due simply to idiosyncratic lineage sorting can also characterize taxa that separated anciently but whose speciation events were close in evolutionary time (Fig. 3). In such cases, the lineages from the polymorphic ancestral gene pool that happen to have reached fixation in distant descendants are those that produced the original gene-tree/species-tree disharmony (Takahata, 1989; Wu, 1991). For example, with respect to organismal phylogeny, taxa D and E in Fig. 3 are members of a clade to the exclusion of F, whereas with respect to the gene tree in Fig. 3, taxa E and F are members of a clade to the exclusion of D. Thus, in principle the gene-tree/species-tree "problem" is not confined to recently separated taxa.

EXAMPLE

The overall phylogeny for human, chimpanzee, and gorilla appears to be a near trichotomy with the most

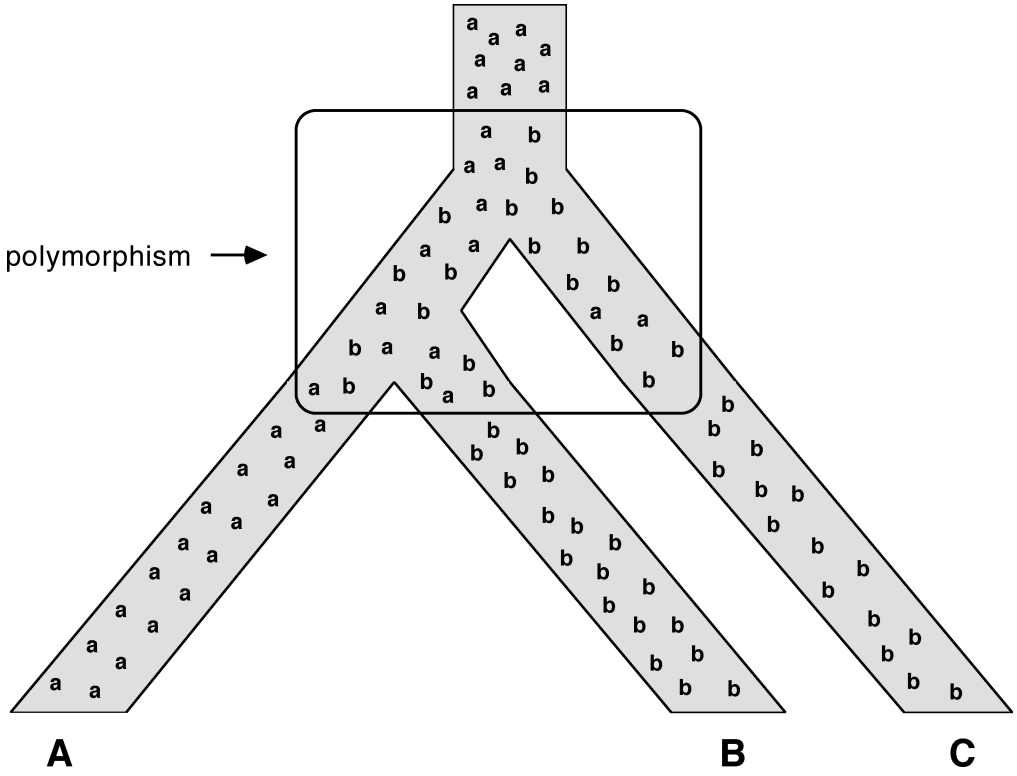

FIGURE 2. Another depiction of how alleles at a gene (or alternative states of any polymorphic trait) can be misleading with respect to the tree topology for the species in which the alleles are housed. Shown is a polymorphism that traversed successive speciation nodes only to sort idiosyncratically and later become fixed in the descendant species in a pattern that at face value would appear to be discordant with the species phylogeny.

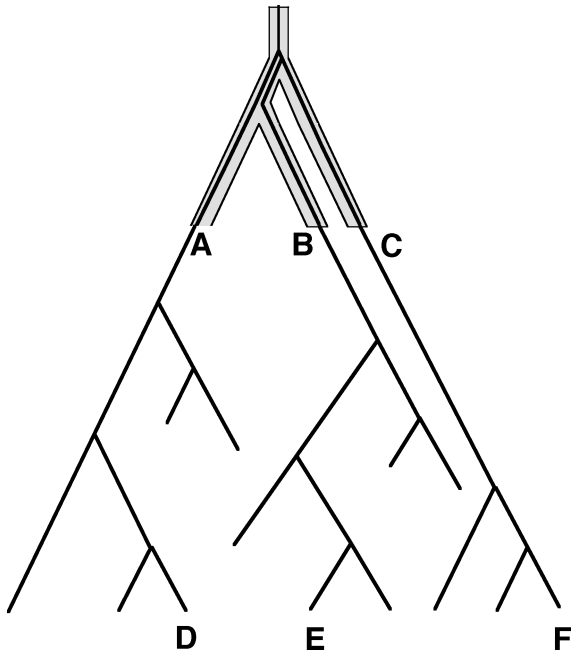

FIGURE 3. Diagrammatic representation of how an ancient discordance between a gene tree and a species tree can be perpetuated indefinitely and thereby retained as a permanent incongruity between the gene tree and the species tree of descendant taxa (after Avise, 2000).

likely resolution being sister-taxa status for *Homo* and *Pan* to the exclusion of *Gorilla* (e.g., Stanyon et al., 2006). Not all homoplasy-free gene trees or sets of character states are expected to match this composite species topology, however, if some polymorphisms happen to have traversed the adjacent evolutionary nodes before sorting idiosyncratically into various pairs of the descendant taxa (Takahata et al., 1995). For example, Chen and Li (2001) reported that whereas DNA sequences from each of 31 independent loci support the *Homo-Pan* clade, 12 appear to support a *Pan-Gorilla* clade and 10 appear to support a *Homo-Gorilla* clade; and in a more extensive recent analysis, Ebersberger et al. (2007) reported that about 23% of 23,210 DNA sequence alignments in the great apes implied at face value that chimpanzees are not the closest genetic relatives of humans (see Patterson et al., 2006, for comparable findings based on 20 million base pairs of aligned human and chimpanzee sequence). In the human-chimpanzee case, the causes of these discrepancies are not fully understood (and may include postspeciation introgression; Patterson et al., 2006). However, at least in principle, each gene tree could be correct in the sense of providing valid genealogical signal (i.e., without homoplasy) for the specific portion of the genome that it represents.

DEFINITION

A formal term seems desirable to encapsulate the essence of the phylogenetic processes described above that can lead to genuine discordances between particular gene trees (components of the genome) and a composite or overall species phylogeny. We suggest the word *hemiplasy*, because the responsible lineage sorting processes have homoplasy-like consequences despite the fact that the character states themselves are genuinely homologous and apomorphic. So, hemiplasy is somewhat like homoplasy in terms of its face-value phylogenetic consequences, yet its evolutionary etiology is fully distinct from homoplasy. We suggest the following formal definition of hemiplasy: the topological discordance between a gene tree and a species tree attributable to lineage sorting of genetic polymorphisms that were retained across successive nodes in a species tree. A set of hemiplasious alleles, genes, or other character states would thus be those that contribute to hemiplasy in a phylogenetic data set.

Other evolutionary processes are also capable of producing homoplasy-free discordances between gene trees and species phylogenies. For example, particular alleles can leak across species boundaries via hybridization and introgression, and pieces of DNA sometimes move between species via true horizontal transfer (viral-mediated, for example). In such cases, the transferred DNA is a bona fide part of the genetic history of the species in question, but the gene tree would differ dramatically from the majority phylogeny for the remainder of the genome. We acknowledge that such outcomes may be difficult to distinguish from genuine hemiplasy in some particular empirical instances. Nevertheless, for epistemological clarity we recommend that the term hemiplasy not include these additional (and well appreciated) generators of phylogenetic discordance between gene trees and species trees but instead be confined to discordances that arise from idiosyncratic lineage sorting per se.

The importance of the hemiplasy concept is further evidenced by the fact that several recently introduced phylogenetic approaches in effect acknowledge and attempt to accommodate the phenomenon (Carstens and Knowles, 2007; Edwards et al., 2007; Liu and Pearl, 2007; Maddison and Knowles, 2006). Additional cutting-edge research of this type would likely be stimulated and more widely appreciated if a simple term (hemiplasy) were available to replace the cumbersome phraseologies in current use.

Hemiplasy is not a synonym for lineage sorting; rather, it is a consequence of lineage sorting. Furthermore, hemiplasy is not the exclusive outcome of lineage sorting; indeed, it is usually a minority outcome compared to the larger number of hemiplasy-free gene trees that normally are expected to comprise a typical species tree.

PEDAGOGICAL RATIONALES

Many words in the extensive lexicon of systematics encapsulate conceptually challenging phylogenetic notions. Yet these terms have been widely adopted, much to the benefit of the systematics community. For example, students typically gain access to a discipline by first learning its language and definitions, and the mere act of formally naming and distinguishing subtle notions (such as the concept of a synapomorphy versus a symplesiomorphy or of paralogy versus orthology) can have obvious pedagogical advantages. Seasoned professionals also benefit from having formal terms available that enable them to discuss complex topics with streamlined language.

If the term hemiplasy is adopted and employed widely, it will undoubtedly help to (1) foster thought and discussion on the ineluctable but oft-neglected phylogenetic ramifications of lineage sorting within and among related populations and species; (2) promote the fundamentally important conceptual distinction between a gene tree and a species tree, including the notion that any cladogram for a group of organisms is really a statistical "cloudogram" of gene trees with a variance (Maddison, 1997); (3) disabuse systematics of the longstanding but invalid notion that even a single synapomorphy is sufficient for the recognition of an organismal clade in sexually reproducing taxa; (4) promote the routine incorporation of information from multiple unlinked genes or other independent characters into phylogenetic reconstructions at the levels of populations, species, and higher taxa; and (5) foster searches for categories of genetic markers that not only are likely to be homoplasy free but also that are less prone to hemiplasy. Markers that should be relatively immune to hemiplasy might include, for example, those with smaller effective population sizes (such as cytoplasmic loci compared to autosomal nuclear genes) or those that are likely to experience underdominant selection (such as some types of chromosomal markers).

In conclusion, adoption of the word hemiplasy should contribute to the injection of oft-neglected "population thinking" into phylogenetic assessments and thereby provide another incremental step toward unifying the traditionally disparate fields of phylogenetic biology and population genetics.

REFERENCES

Avise, J. C. 2000. Phylogeography: The history and formation of species. Harvard University Press, Cambridge, Massachusetts.

Avise, J. C. 2004. Molecular markers, natural history, and evolution, 2nd edition. Sinauer Associates, Sunderland, Massachusetts.

Carstens, B. C., and L. L. Knowles. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: An example from *Melanopus*. Syst. Biol. 56:400–411.

Chen, F. C., and W. H. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Human Genet. 68:444–456.

Degnan, J. H., and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:762–768.

Doyle, J. J. 1992. Gene trees and species trees: Molecular systematics as one-character taxonomy. Syst. Bot. 17:144–163.

Ebersberger, I., P. Galgoczy, S. Taudien, S. Taenzer, M. Platzer, and A. von Haeseler. 2007. Mapping human genetic ancestry. Mol. Biol. Evol. 24:2266–2276.

Edwards, S. V., L. Liu, and D. K. Pearl. 2007. High resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA 104:5936–5941.

Hennig, W. 1950. Grundzüge Einer Theorie der Phylogenetischen Systematik. Deutscher Zentralverlag, Berlin. [Published in English translation in 1966: Phylogenetic systematics, University of Illinois Press, Urbana, Illinois.]

Hudson, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution 37:203–217.

Liu, L., and D. K. Pearl. 2007. Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504–514.

Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Maddison, W. P., and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

Neigel, J. E. and J. C. Avise. 1986. Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. Pages 515–534 *in* Evolutionary processes and theory (S. Karlin and E. Nevo, eds.). Academic Press, Orlando, Florida.

Nichols, R. 2001. Gene trees and species trees are not the same. Trends Ecol. Evol. 16:358–364.

Patterson, N., D. J. Richter, S. Gnerre, E. S. Lander, and D. Reich. 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature 441:1103–1108.

Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and species trees. Theor. Pop. Biol. 61:225–247.

Stanyon, R., D. Caramelli, and B. Chiarelli. 2006. Molecular views of human origins. Human Evol. 21:19–31.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.

Takahata, N. 1989. Gene genealogy in three related populations: Consistency probability between gene and population trees. Genetics 122:957–966.

Takahata, N., and M. Nei. 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110:325–344.

Takahata, N., Y. Satta, and J. Klein. 1995. Divergence time and population size in the lineage leading to modern humans. Theor. Pop. Biol. 48:198–221.

Wu, C.-I. 1991. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. Genetics 127:429–435.

First submitted 15 October 2007; reviews returned 25 January 2008; final acceptance 29 February 2008

Associate Editor: Laura Kubatko